An Analysis of Trends in Movies

October 7, 2010

1 Introduction

While the film industry has recently exploded, movie-going has long been a common pastime in America. Because of this there are hundreds of thousands of movies, some of which are loved, some of which are hated. Some genres spark more love or hatred than others. In our analysis we investigate trends in popularity of specific genres of movies. Our definition of popularity varies to mean popularity in production, popularity in number of votes, or popularity by overall rating over time. By looking at popularity from various angles, we are able to investigate effects of a movie's runtime on its popularity, possible causes of films that polarize ratings, and how America has tended to perceive specific genres of film over the years.

2 Data Cleaning

From our explorations in previous weeks, we realized that the budget column was responsible for all of the missing values in our data set. Because we are not specifically investigating budget in our analysis and simply looking at the movies with available budget data would greatly bias our data, we created a separate data frame without the budget column. Then we set about categorizing our data in terms of its genre. While the original data had a binary system of identifying a particular movie's genre, in order to better classify, sort, and label subsets, we established a new variable, genre. This variable labels all movies with a single genre with that genre and also pulls out movies that belong to particular popular mixed genres: Animated comedy, Romantic comedy, Docu-drama, and Animated short. Finally, we decided to look only at movies that had a more 'traditional' runtime of under 300 minutes in order to avoid distortions of the data by the few extremes.

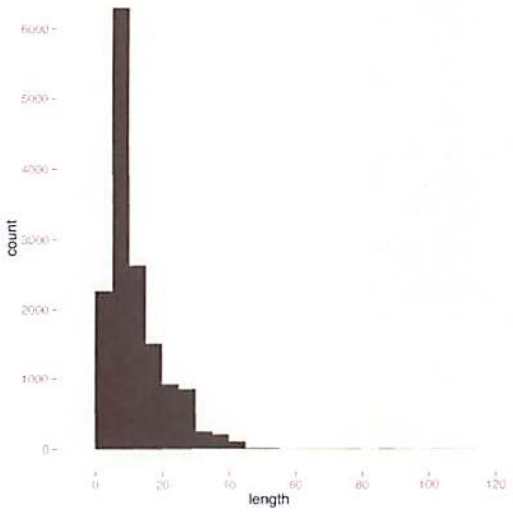
3 Correlations with Length

3.1 What is a Short Movie?

First and foremost, what is a short movie? The Academy of Motion Pictures Arts and Sciences (AMPAS) defines a short movie as one under 40 minutes. [1] However, as shown in figure 1 it appears that our data defines short movies as under 50 minutes, as evidenced by the following histogram of lengths of short movies.

However, it also shows an anomaly of short movies near the 90 minute mark. These movies belong to an average of more than 2 genres, which may imply that belonging to other genres lengthens a movie's runtime, but because these entries are so far from the given definition for a short movie, these outliers are most likely due to a data entry error and will be removed from following calculations.

[1] http://www.oscars.org/awards/academyawards/rules/rule19.html



great caption :)

Figure 1: A histogram of the lengths of short movies with bin width of 5. It is right skewed and centered around x = 10. We also see clusters of potential outliers out at the tail.

3.2 Variations in Length in Short Films

Keeping our focus on short movies, there also appears to be a trend in the runtimes of different types of short film. This figure 2 displays the overall density of the lengths of the two subgenres. The sharp spike in short animated films suggests a small standard deviation, which we find to be 5.27, almost half that of the standard deviation of solely short films.

In questioning the source of this contrast, one may look at the years of release of these short films. Figure 3 shows that live action shorts have become popular relatively recently, while animated shorts saw a jump in releases in the 1930s. AMC reports [2] that Disney released its first animated short in the 1932 which marked the start of colored films. It is possible that the lack of variation in the runtimes of animated films is due to the monopolization of such films by companies such and Disney and Pixar, who began its foray into animated shorts in the mid 1980s, [3] though the data would have to be updated with the addition of the name of the releasing studio to be able to investigate this theory.

good hypothesis

3.3 Length and Ratings

With the sudden explosion in the popularity of short films recently, one may wonder if certain lengths are more likely to be rated more highly, or else, if certain ratings lend themselves to a general length. We can investigate these trends with figure 4 in short films and find that short films around the average rating of 6 tend to have a slightly smaller and less varied runtime.

Expanding this question out to focus on all genres, a similar plot, figure 5, of ratings vs. length shows that the standard deviation in lengths for the highest rated movies are much larger than those of lower ranked movies. Presumably this is due to the fact that many more factors influence a movie's rating than just its length, as we will now discuss.

[2] http://www.filmsite.org/30sintro.html
[3] www.pixar.com/shorts/index.html

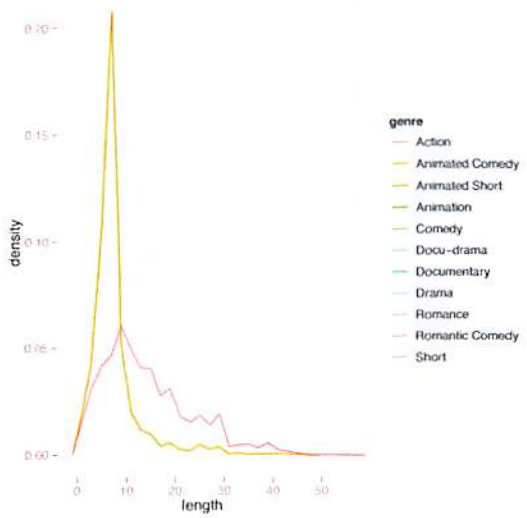


Figure 2: A density plot of the lengths of short and animated short movies with bin width of 2. Evident in this graph is the very tight cluster of animated short lengths compared to the varied lengths of live action shorts.

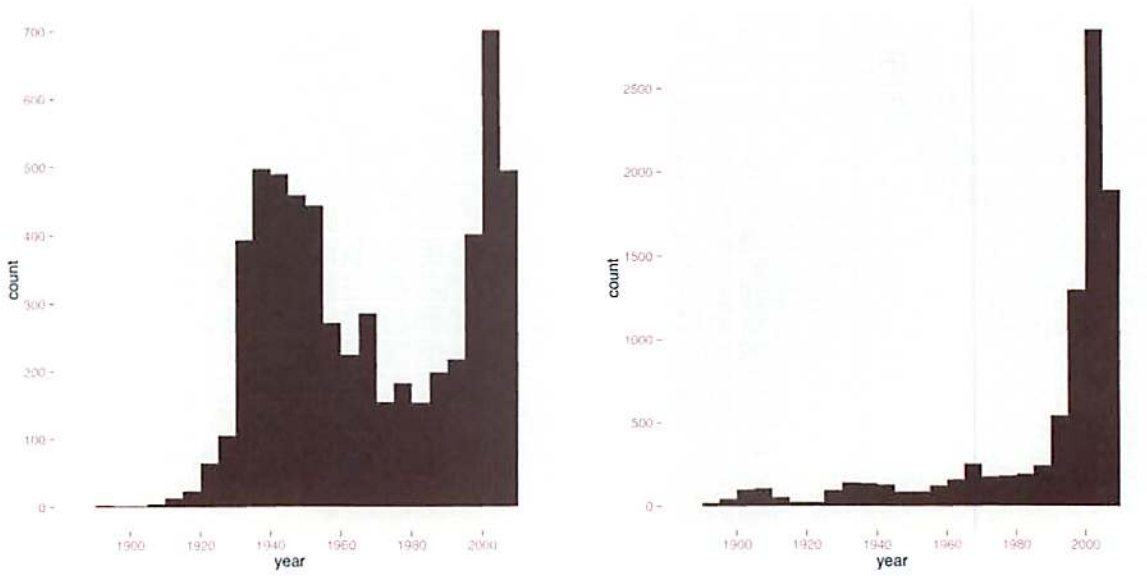


Figure 3: (Left) A histogram of the year of release for animated short films with bin width of 5. It is bimodal, with notable spikes occur in the 1930s and 1990s. (Right) A histogram showing the release year for live shorts, with bin width 5. This distribution is left skewed, with only a spike around 1990.

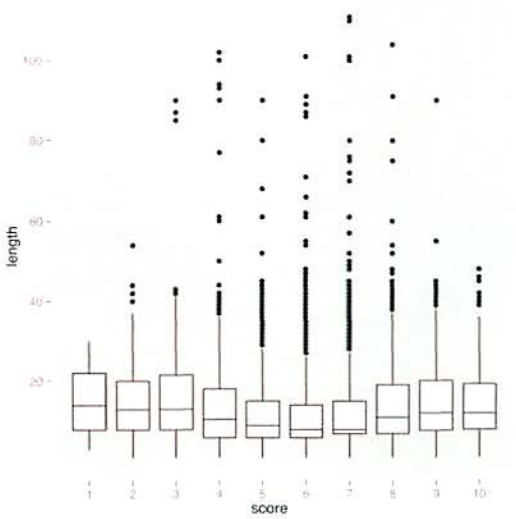


Figure 4: Boxplots of length by the rounded rating score for short films. There is a slight dip in the median length as scores get closer to the average.

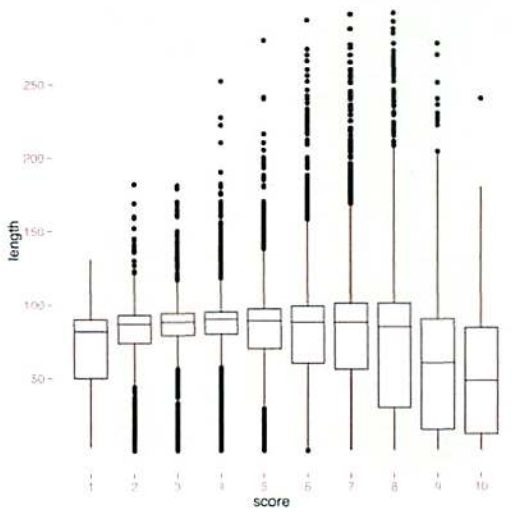


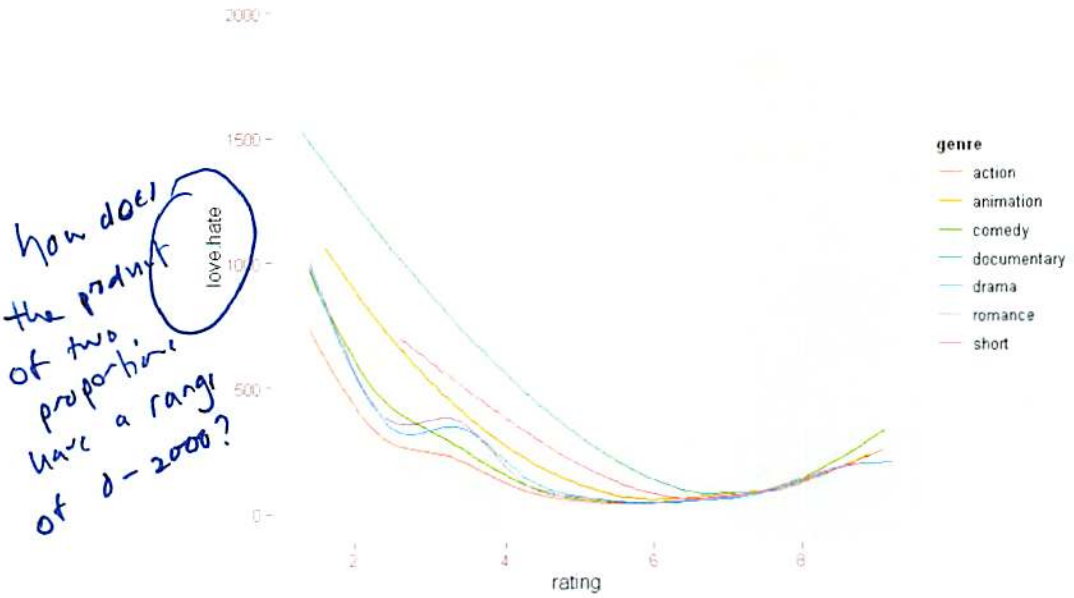
Figure 5: Boxplots of length by the rounded rating score for all films. As ratings increase, the median length seems to decrease while the standard deviation greatly increases.

4 Measuring Polarizing Opinions of Movies

The goal of this plot was to investigate the properties that contribute to a polarizing movie. We are defining a polarizing movie as one that has a larger percentage of very low and very high ratings.

To do this, we created a new variable called "love.hate", which was equal to the proportion of votes rating the movie as a one multiplied by the proportion of movies rated as tens. This multiplier gives a reasonably accurate representation of the degree of polarity for movie ratings. — can you provide evidence for this?

We then graphed love.hate against the average rating to look at any possible relationships



how does the product of two proportions have a range of 0 - 2000?

Figure 6: Polarizing Rating against Average Rating by Genre

The graph shows a very negative non-linear relationship between love.hate and rating. Once can also see that movies that tended to have an overall poor reception by the public occasionally still had a "cult following." However, movies that were generally popular did not have a large proportion of people who rated the movie very poorly, that is, the results were less polarizing for more popular movies.

We found it interesting that documentaries were by far the most polarizing genre. Since this genre includes a variety of programs, such as educational videos and concerts, it seems that, unless people were really interested in the subject or great fans of the band, the reviews were terrible. Likewise, fans of the band or those interested in the educational topic really seemed to enjoy the program.

Action was the least polarizing genre, which does seem to make sense. Since action films tend to have less complicated plots and rely more so on visual effects to entertain the audience. Therefore it stands to reason that they tend to be less polarizing as long as they have their redeeming quality of intense scenes and visual effects.

good hypotheses!

5 Distribution of Ratings by Genre

Next, we examined the relationship of the distributions of ratings amongst the various genres. We thought that possibly variations in the distributions existed amongst the various genres. For instance, we hypothesized that, since action movies tend to receive less critical reviews, action would have a less polarizing genre than the other distributions.

To create this plot we created vectors of the mean proportion for each rating, one through ten, for each genre. We then plotted these different means against the possible ratings.

We also included the distribution of the upper echelon of love.hate ratings as detailed in the last section. This was to compare how the polarizing distributions compare to the average distribution of ratings. We also included the distribution of ratings for the most polarizing movie, which was *Jonas Brothers: The 3D Concert Experience*.

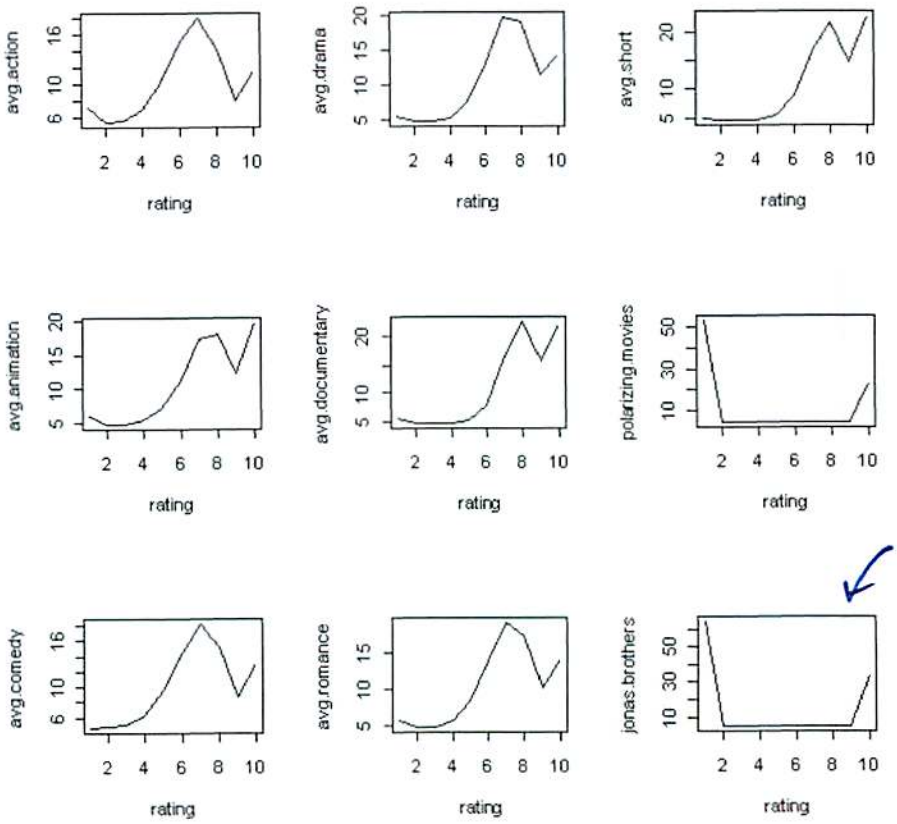


Figure 7: Average Distribution of Ratings by Genre, and Distribution of Polarizing Movies.

The average distributions of rating for each genre were reasonably similar. Our hypothesis that action would perhaps be less polarizing, because of the genres emphasis on special effects rather than plot, does seem to have some weight. Action had the widest middle peak, and had a lower mean proportion of voters choosing 10. However, action had a relatively high proportion of one votes.

The two plots on the lower right showing the distribution of movies with polarizing opinions were dras-

tically different than the average distributions. The peaks were at the two tails of the distributions, however this was to be expected as the subset was sorted based on this criteria.

Looking specifically at the Jonas Brother's Concert, our feeling is that people entered the film with a bias that influenced their ratings. People who were fans of the Jonas Brothers before watching the film would leave the movie with a positive review of the film simply because they liked the music. However, if the actual film was a flop as shown by their average rating below 2, a large proportion of people who were not particularly fans of the band tended to be very critical of the film.

great observation

These plots also showed another bias. Every plot of the average distribution by genre showed a significant drop in the proportion who voted for a rating of 9. This led us to believe that perhaps having a rating system from 1 to 10 is not adequate, because people seem to arbitrarily determine a difference between voting for nine and voting for 8 or 10. The next section addresses this issue, grouping together ranges of ratings into a grade, A, B, C, etc., fixed by equal percentage.

the section after next?

6 Time Series Analysis: Genre Production and Characteristics

Examining the evolution of various characteristics such as viewer ratings, using the IMDB viewer data, production numbers for each genre, and even MPAA ratings over time show interesting trends of the data and genre categories. It in a sense gives a window to examine and analyze changing trends in popularity and production. We begin by showing the most general plot, a simple time series of production of movies for each genre classification.

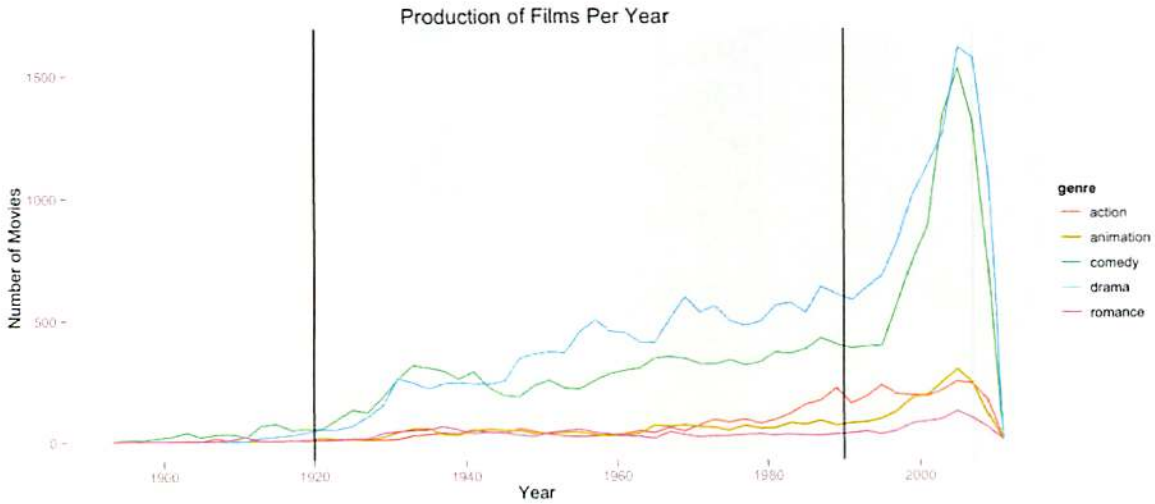


Figure 8: Time Series showing production numbers of films per year, including movies of length 60 min. to 180 min., and separated for each genre of interest.

In showing this plot, we wish to make a special note that IMDB was founded in 1990. [4] The influence of IMDB has to be specially considered in offering any analysis over time. This will be discussed later. There is an obvious trend for movies to increase their production in the long run, specifically referencing the Drama and Comedy genres, which show the largest and most pronounced increase in production.

[4] http://www.imdb.com/features/anniversary/2010/ discusses their 20th anniversary and also asserts their copyright has been maintained from 1990 to 2010.

In the plot, we also show three special intervals:

- 1990 to 1920
- 1920 to 1990
- 1990 to 2010

From 1900 to 1920, no one genre had a significantly differentiated production from any other genre. For some, there is an incredibly constant trend, drama and comedy vary slightly. It is worth observing that the production of any one year is rather low, especially when compared with the later years.

From 1920 to 1990, the largest time interval, we see the divergence of the genres comedy and drama, which sharply increase their production per year after 1920. This trend maintains throughout the interval. An additional note, the year 1960 marks the divergence of the other three genre, romance, animation, and action. Whereas, those genre previously remained approximately constant with small fluctuations, the genre production per year diverge here.

Finally, from 1990 to 2010, the present, there is an incredibly sharp increase in production per year of comedy and drama. The other genre approximately see their production unchanged from the increases that occurred previously occurred before 1990.

6.1 Comments about the creation of IMDB

It is worth noting that the institution of IMDB would likely occur at a time when both the internet and movie popularity were becoming more popular and accessible. This intuition is reflected in the outcome of production increases from 1990 onward. The output of the movie industry greatly increased, and as we will show the viewer responses also increases from about 1990 onward. This will be important to consider in examining how the institution of IMDD, specifically the year it was founded, affects the voter turnout. Since, we have shown that voter response increased from 1990 onward, it has been shown that low voter trends can greatly affect the ratings of genre and potentially lead to incorrect or skewed distributions. This creates a barrier in examining and analyzing how ratings evolve over years for various genre. But, it does not prohibit analysis and intuition.

7 Evolution of ratings over time

We begin by showing a histogram demonstrating the production aspects of the first plot, but incorporating a new feature, a fill effect to display the grade system. A brief note about the grade system. It was created specifically for this plot as a way to categorically label a particular movie. The algorithm applied is as follows:

- A Rating (7.2, 10]
- B Rating (6.5, 7.2]
- C Rating (5.9, 6.5]
- D Rating (4.9, 5.9]
- F Rating [1, 4.9]

We calculated these levels ~~using the `cut_number` command, requesting that the ratings be split by percentage into 5 groups corresponding to a letter grade~~. This was to allow comparison of the relative change over time that each genre and grade exhibited. We then show the proportion of ratings or grade received for each year. It is important to note that this has been normalized, see appendix for details and code.

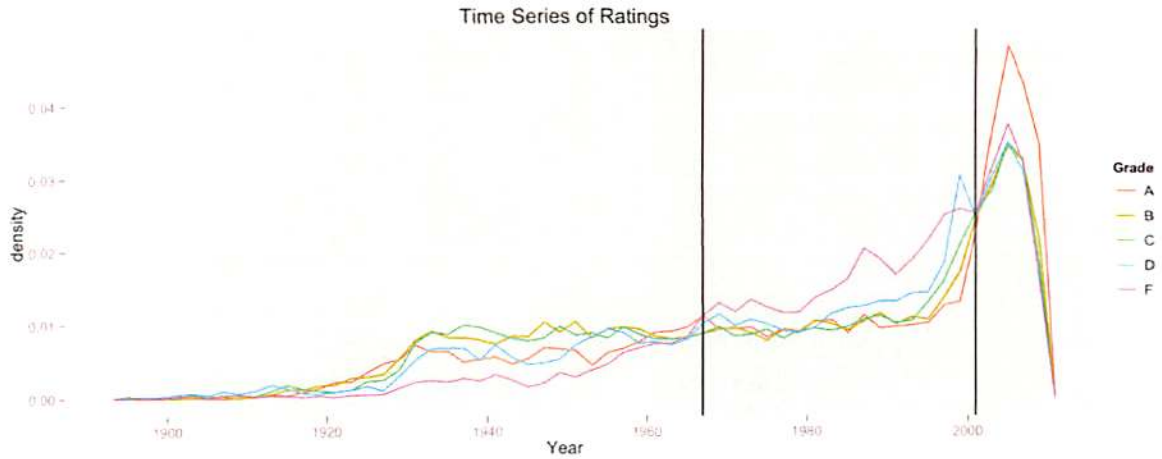


Figure 9: Time series showing density of films per year produced that received a specified grade to achieve an equal distribution over the entire data set of interest. Bin width equal to 2.

Examining the year 2001 onward, there is a distinct trend for more movies to be graded an A relative to the other grades, which appear to follow the same trend line. This is sharply contrasted by the preceding time interval, which we show in between the two vertical lines. Here, the grade of F dominates, showing that, relatively, in this time interval more movies were given a rating of F than any other. To examine how this could necessarily be the case, we offer the same plot faceted by genre.

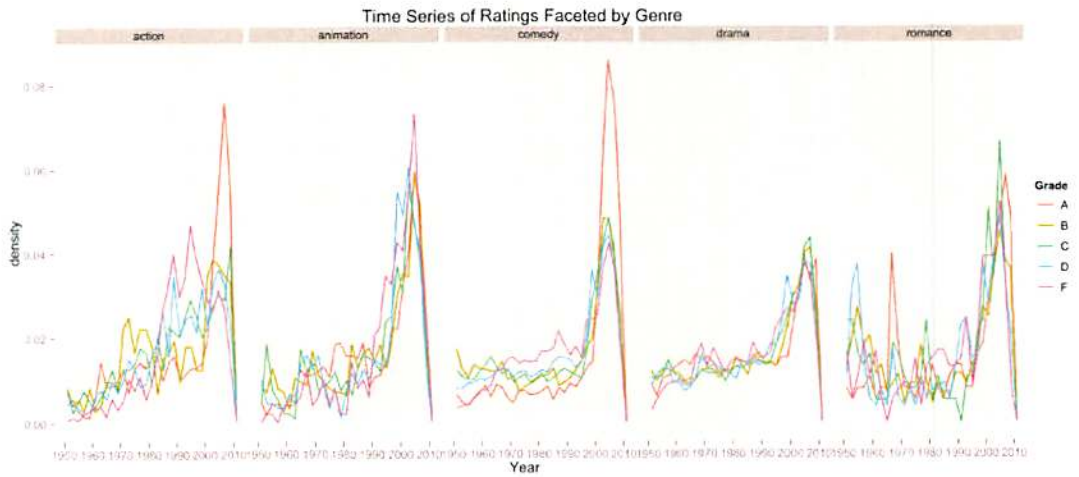


Figure 10: The time series above, except now faceted, shows the same information as the previous figure for each genre specifically, showing from 1950 to 2010. Bin width equal to 2.

It might not be immediately apparent, but one genre in particular has the same form on the intervals we wished to consider. Specifically, comedy has the same general form as the preceding plot. Most importantly from 2001 onward the grade A assigned to comedy movies greatly outweighed all other grades, relatively. By examining again the first figure, we see that comedy on this time interval had a considerable amount

and action?

of movies, very many more than action, romance, and animation, and approximately the same (actually a little less on the average) as drama. This magnitude of movies with, as demonstrated, a higher proportion of grade A movies from 2010 onward, very well affects the total relative measure of the grades and offers insight into this phenomenon that not all movies necessarily received a grade of A, but rather, one genre that by production numbers dominated most other genre featured this phenomena.

One last consideration needs to be addressed, the number of votes per year. The respondence of votes for movies in each year should be examined since examine how the ratings have changed over time. The hypothesis is that as we see more movies increasing over time, we should as well observe more votes each year since there are more movies for which one can vote. This is easy to show with the following figure.

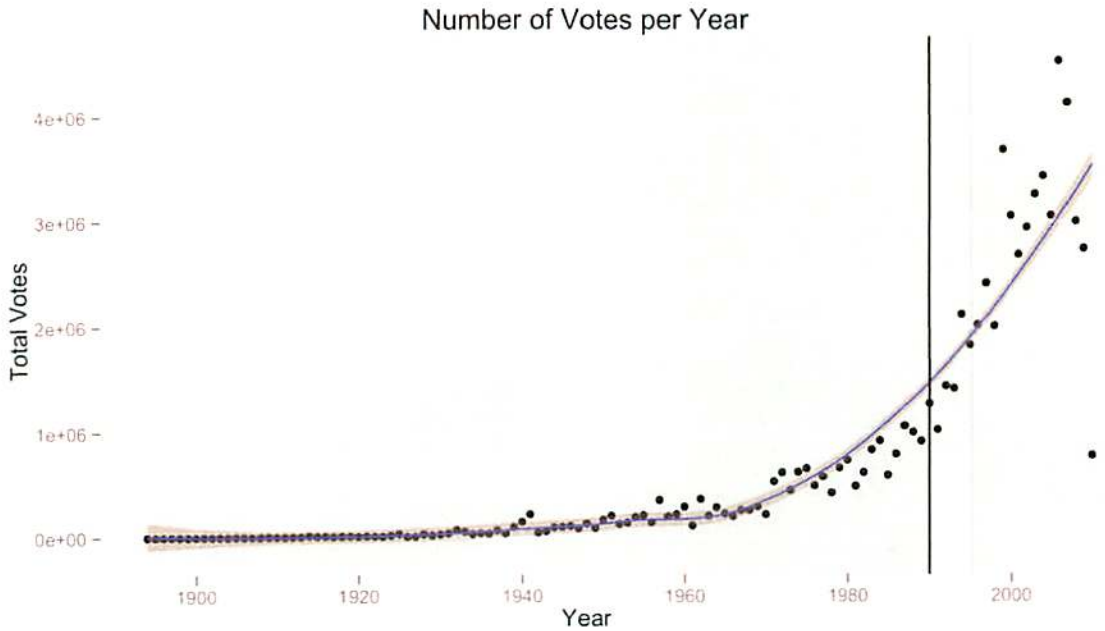


Figure 11: The plot above shows the number of votes for all the movies of interest in each particular year. A smooth curve has been added to show the positive relationship.

We have highlighted the year 1990 for reference as to when IMDB was created. This is also to show that the increasing trend began before that year and is not necessarily marked with the creation of IMDB. This implies that the votes per each year are as much a function of those around to see them and with heavy tribute to the increasing production of those movies over time.

8 Conclusion

With the multitude of movies, quantifying the different characteristics and genre provides many opportunities for analysis. We have found that the popularity of movies can be correlated with unexpected attributes, such as the length of a short film. Intriguing questions like these beg for insight and explanation into what does cause or effect ratings skews other than public opinion. If public opinion is very polarized or divergent, then the rating can be low, still with what was dubbed a "cult following." But, we did not observe the converse. Showing the distributions, we further illustrate this point and argue one of the potential pitfalls of analyzing movies on the average, specifically with something as subjective as rating and public opinion.

Movies by nature are incredibly differentiated products and we show how this is apparent within the distribution of votes for subsets of each genre, referring to the love-hate skew definition we provided. Time also is a major player. Over the years, popularity of different genre vary, grow and fall. We have observed various time intervals in which unique behavior occurs, showing the positive correlation with number of movies and relative ranking, with respect shown to the number of votes cast in each year. The industry is a complicated field, and analysis involves many endogenous variables which make it difcult to know with certainty if length of movies does affect the ratings of this group of movies or if responsibility should be assigned to a voting skew, or an increased growth trend. While the latter seems more probable, more data in the form of hard numbers, release months, and country of origin would be helpful forthese questions to be examined and to engage in investigation of one of Americas biggest industries.

P 4
C 4
E 4

A Code

A.1 Data Cleaning

```
library(ggplot2)
movies_2 <- read.csv(file = "movies.csv.bz2")

# need to remove NA first --------------------
# NAs only in budget data so can remove budget column
no_budget <- movies_2
no_budget$budget <- NULL

# or can just remove NA budget movies, but this removes
    # quite a chunk of our data unnecessarily
clean <- na.omit(movies_2)

# finding single category movies------------
cat <- rep(NA, nrow(no_budget))
for (i in 1:nrow(no_budget)){
  cat[i] <- sum(no_budget[i, 17:23]) == 1
}

# saving data to avoid repeat calculations
write.csv(cat, "category-vector.csv", row.names = F)
# if recalling use
cat <- read.csv(file = "category-vector.csv")
cat <- cat$x
one_cat <- no_budget[cat, ]

# Labeling single category movies: separate
action <- subset(one_cat, Action == 1)
action$genre <- "Action"
animation <- subset(one_cat, Animation == 1)
animation$genre <- "Animation"
comedy <- subset(one_cat, Comedy == 1)
comedy$genre <- "Comedy"
drama <- subset(one_cat, Drama == 1)
drama$genre <- "Drama"
documentary <- subset(one_cat, Documentary == 1)
documentary$genre <- "Documentary"
romance <- subset(one_cat, Romance == 1)
romance$genre <- "Romance"
short <- subset(one_cat, Short == 1)
short$genre <- "Short"

# And rejoin
one_genre <- action
one_genre <- join(one_genre, animation, type = "full")
one_genre <- join(one_genre, comedy, type = "full")
```

could basically omit this

also check out the rowSums functio

by = ?

but I think you probably want rbind here.

```
one_genre <- join(one_genre, drama, type = "full")
one_genre <- join(one_genre, documentary, type = "full")
one_genre <- join(one_genre, romance, type = "full")
one_genre <- join(one_genre, short, type = "full")
one_genre <- one_genre[order(one_genre$title), ]

# Finding double category movies ----------------
multi_cat <- no_budget[!cat, ]

# Labeling common multi genred movies: separate
short_ani <- subset(multi_cat, Animation == 1 & Short == 1)
short_ani$genre <- "Animated Short"
ani_comedy <- subset(multi_cat, Animation == 1 & Comedy == 1)
ani_comedy$genre <- "Animated Comedy"
docu_drama <- subset(multi_cat, Documentary == 1 & Drama == 1)
docu_drama$genre <- "Docu-drama"
roman_comedy <- subset(multi_cat, Romance == 1 & Comedy == 1)
roman_comedy$genre <- "Romantic Comedy"

# And rejoing
multi_genres <- short_ani
multi_genres <- join(multi_genres, ani_comedy, type = "full")
multi_genres <- join(multi_genres, docu_drama, type = "full")
multi_genres <- join(multi_genres, roman_comedy, type = "full")

# Forming the full set ---------------------------
common_genres <- join(multi_genres, one_genre, type = "full")
common_genres <- common_genres[order(common_genres$title), ]
# want to look at movies without extreme runtimes
common_genres <- subset(common_genres, length < 300)
write.csv(common_genres, "common-genres.csv", row.names = F)
```

A.2 Section 3: Length of Various Genre

```
# Length of Short Films ----------------------------------

# What seems to be the limit for a short movie?
short <- subset(common_genres, Short == 1)
qplot(length, data = short, binwidth = 5)
ggsave(file = "short_length_hist2.pdf")
# also tried bin width of 10 and 2
# 2 shows under 50 minutes tends to be the consensus, but film at 90?
long_short <- subset(short, length > 60)
nrow(long_short)

# 41 films over 60, multiple genres may influence length:

n_genres <- rep(NA, nrow(long_short))
for (i in 1:nrow(long_short)) {
```

```
    n_genres[i] <- sum(long_short[i,17:23])
}
mean(n_genres)
# 2.170732


# Length variations in various short films --------------------------

short <- subset(short, length < 60)
qplot(length, ..density.., data = short, binwidth = 2,
    color = genre, geom = "freqpoly")
# Can see much smaller variation in animated short movies, so remove extra
# genres for clarity
subset(common_genres, genre == "Short" | genre == "Animated Short")
qplot(length, ..density.., data = short, binwidth = 2,
    color = genre, geom = "freqpoly")
# Also tried binwidth = 5
ggsave(file = "short_length_poly.pdf")


# Actual standard deviations reinforce this point
sd(short[short$genre == "Short", ]$length)
# 9.173726
sd(short[short$genre == "Animated Short", ]$length)
# 5.265871

# Does the release year offer explanations for this difference?
ani_short <- subset(short, genre == "Animated Short")
qplot(year, data = ani_short, binwidth = 5)
ggsave(file = "ani_short_year.pdf")
live_short <- subset(short, genre != "Animated Short")
qplot(year, data = live_short, binwidth = 5)
ggsave(file = "live_short_year.pdf")


# Length versus rating ----------------------------------

# Find average score for short movies and for all movies
mean(short$rating)
# 6.501258
mean(common_genres$rating)
# 6.176568
# Round ratings into interger groups
scored_short <- ddply(short, c("title", "rating", "length", "genre"),
    summarise, score = as.factor(round(rating)), .progress = "text")
# Do different ratings come from different length movies
qplot(reorder(score, rating), length, data = scored_short,
    geom = "boxplot") + xlab("score")
```

```
ggsave(file = "short_scores.pdf")

# Also explore with full set of movies
scored_common <- ddply(common_genres, c("title", "rating", "genre", "length"), summarise,
    score = round(rating), .progress = "text")
scored_common$score <- as.factor(scored_common$score)
qplot(score, length, data = scored_common, geom = "boxplot")
ggsave(file = "all_scores.pdf")
# Can see a parabolic like trend
```

A.3 Section 4: Polarizing Ratings

```
movies <- read.csv("movies.csv.bz2")

# This creates my measure for a movies polarizing opinions
movies$love.hate <- movies$r1 * movies$r10

# Focusing only on movies with a large number of votes
popular <- subset(movies, votes > 1000)

# The next group of lines subset to view relationships by genre
action <- subset(popular, Action == 1)
action$genre <- print("action")

animation <- subset(popular, Animation == 1)
animation$genre <- print("animation")

comedy <- subset(popular, Comedy == 1)
comedy$genre <- print("comedy")

drama <- subset(popular, Drama == 1)
drama$genre <- print("drama")

documentary <- subset(popular, Documentary == 1)
documentary$genre <- print("documentary")

romance <- subset(popular, Romance == 1)
romance$genre <- print("romance")

short <- subset(popular, Short == 1)
short$genre <- print("short")

# Recreates the original movie matrix with a new column of genre, I chose to just have movies
# with multiple genres just listed multiple times with each line having their various genres
popular.movies <- rbind(action, animation, comedy, drama, documentary, romance, short)

qplot(rating, love.hate, data = popular.movies, geom = "smooth", color = genre)
```

didn't you already create these?

```
# The next couple lines were just to get an idea of the most polarizing movies, I did not include
# the plots in the report, they were just purely for my purposes of talking aobut the data
love.hate.1500 <- subset(popular.movies, love.hate > 750 & year > 2004)

qplot(rating, love.hate, data= love.hate.1500, color = genre, geom = "jitter")
```

A.4 Section 5: Distribution of Ratings by Genre

```
# Code applied to the next section over Distribution of
# Ratings by Genre and the distributions for polarizing
# movies to compare any possible differences.
polarizing.movies <- c(mean(love.hate.1500$r1),
  mean(love.hate.1500$r2),
  mean(love.hate.1500$r3),
  mean(love.hate.1500$r4),
  mean(love.hate.1500$r5),
  mean(love.hate.1500$r6),
  mean(love.hate.1500$r7),
  mean(love.hate.1500$r8),
  mean(love.hate.1500$r9),
  mean(love.hate.1500$r10))

# I also wanted to look specifically at the most
# polarizing movie.
jb.concert <- subset(love.hate.1500, love.hate > 2000)

jonas.brothers <- c(jb.concert$r1, jb.concert$r2,
  jb.concert$r3, jb.concert$r4, jb.concert$r5,
  jb.concert$r6, jb.concert$r7, jb.concert$r8,
  jb.concert$r9, jb.concert$r10)

highest.love.hate.rating<- c(mean(love.hate.1500$r1),
  mean(love.hate.1500$r2), mean(love.hate.1500$r3),
  mean(love.hate.1500$r4), mean(love.hate.1500$r5),
  mean(love.hate.1500$r6), mean(love.hate.1500$r7),
  mean(love.hate.1500$r8), mean(love.hate.1500$r9),
  mean(love.hate.1500$r10))


avg.action <- c(sum(action$r1)/length(action$r1),
  sum(action$r2)/length(action$r2),
  sum(action$r3)/length(action$r3),
  sum(action$r4)/length(action$r4),
  sum(action$r5)/length(action$r5),
  sum(action$r6)/length(action$r6),
  sum(action$r7)/length(action$r7),
  sum(action$r8)/length(action$r8),
```

```
  sum(action$r9)/length(action$r9),
  sum(action$r10)/length(action$r10))

avg.animation <- c(sum(animation$r1)/length(animation$r1),
  sum(animation$r2)/length(animation$r2),
  sum(animation$r3)/length(animation$r3),
  sum(animation$r4)/length(animation$r4),
  sum(animation$r5)/length(animation$r5),
  sum(animation$r6)/length(animation$r6),
  sum(animation$r7)/length(animation$r7),
  sum(animation$r8)/length(animation$r8),
  sum(animation$r9)/length(animation$r9),
  sum(animation$r10)/length(animation$r10))

avg.drama <- c(sum(drama$r1)/length(drama$r1),
  sum(drama$r2)/length(drama$r2),
  sum(drama$r3)/length(drama$r3),
  sum(drama$r4)/length(drama$r4),
  sum(drama$r5)/length(drama$r5),
  sum(drama$r6)/length(drama$r6),
  sum(drama$r7)/length(drama$r7),
  sum(drama$r8)/length(drama$r8),
  sum(drama$r9)/length(drama$r9),
  sum(drama$r10)/length(drama$r10))

avg.documentary <- c(sum(documentary$r1)/length(animation$r1),
  sum(documentary$r2)/length(documentary$r2),
  sum(documentary$r3)/length(documentary$r3),
  sum(documentary$r4)/length(documentary$r4),
  sum(documentary$r5)/length(documentary$r5),
  sum(documentary$r6)/length(documentary$r6),
  sum(documentary$r7)/length(documentary$r7),
  sum(documentary$r8)/length(documentary$r8),
  sum(documentary$r9)/length(documentary$r9),
  sum(documentary$r10)/length(documentary$r10))

avg.comedy <- c(sum(comedy$r1)/length(drama$r1),
  sum(comedy$r2)/length(comedy$r2),
  sum(comedy$r3)/length(comedy$r3),
  sum(comedy$r4)/length(comedy$r4),
  sum(comedy$r5)/length(comedy$r5),
  sum(comedy$r6)/length(comedy$r6),
  sum(comedy$r7)/length(comedy$r7),
  sum(comedy$r8)/length(comedy$r8),
  sum(comedy$r9)/length(comedy$r9),
  sum(comedy$r10)/length(comedy$r10))

avg.romance <- c(sum(romance$r1)/length(romance$r1),
```

```
  sum(romance$r2)/length(romance$r2),
  sum(romance$r3)/length(romance$r3),
  sum(romance$r4)/length(romance$r4),
  sum(romance$r5)/length(romance$r5),
  sum(romance$r6)/length(romance$r6),
  sum(romance$r7)/length(romance$r7),
  sum(romance$r8)/length(romance$r8),
  sum(romance$r9)/length(romance$r9),
  sum(romance$r10)/length(romance$r10))

avg.short <- c(sum(short$r1)/length(short$r1),
  sum(short$r2)/length(short$r2),
  sum(short$r3)/length(short$r3),
  sum(short$r4)/length(short$r4),
  sum(short$r5)/length(short$r5),
  sum(short$r6)/length(short$r6),
  sum(short$r7)/length(short$r7),
  sum(short$r8)/length(short$r8),
  sum(short$r9)/length(short$r9),
  sum(short$r10)/length(short$r10))


rating <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

# I chose to just create a facet of the different distributions to
# get a less muddled presentation.
par(mfcol=c(3, 3))
plot(rating, avg.action, type = "l")
plot(rating, avg.animation, type = "l")
plot(rating, avg.comedy, type = "l")
plot(rating, avg.drama, type = "l")
plot(rating, avg.documentary, type = "l")
plot(rating, avg.romance, type = "l")
plot(rating, avg.short, type = "l")
plot(rating, polarizing.movies, type = "l")
plot(rating, jonas.brothers, type = "l")
```

A.5 Section 6: Time Series

```
# Method for Setting up Genre variable

movies <- read.csv("movies-mine.csv")
movies <- transform(movies, Genre = 0)

# A general algorithm so that no matter what I wish to
# subset, I can do so with the following templates:
action <- movies$Action == 1 & movies$Animation == 0 &
  movies$Comedy == 0 & movies$Drama == 0 & movies$Romance == 0
```

```
animation <- movies$Action == 0 & movies$Animation == 1 &
  movie$Comedy == 0 & movies$Drama == 0 & movies$Romance == 0

comedy <- movies$Action == 0 & movies$Animation == 0
  & movies$Comedy == 1 & movies$Drama == 0 & movies$Romance == 0

drama <- movies$Action == 0 & movies$Animation == 0 &
  movies$Comedy == 0 & movies$Drama == 1 & movies$Romance == 0

romance <- movies$Action == 0 & movies$Animation == 0 &
  movies$Comedy == 0 & movies$Drama == 0 & movies$Romance == 1

movies$genre[action] <- c("action")
movies$genre[animation] <- c("animation")
movies$genre[comedy] <- c("comedy")
movies$genre[drama] <- c("drama")
movies$genre[romance] <- c("romance")



# Define other logical vectors used for looking up data

length <- movies$length >= 60 & movies$length <= 180
interest <- action | comedy | drama | animation | romance
  & length
```

A.6 Section 7: Evolution of ratings over time

```
# Method for Adding Grade variable:
# Here I wished to create a variable that would give
# each movie a grade, A, B, C, D, E, & F based on
# its rating value

# I wish to assign each grade to an interval that is of
# equal percentage to the others on the entire data set,
# but only the data of interest (defined above)

cut_rating <- cut_number(movies$rating[interest], n = 5)
levels(cut_rating)

# These levels are then taken to be the intervals that we
# assign each respective grade

movies <- transform(movies, Grade = 0)

A <- movies$rating <= 10 & movies$rating > 7.3
B <- movies$rating <= 7.3 & movies$rating > 6.6
C <- movies$rating <= 6.6 & movies$rating > 5.9
D <- movies$rating <= 5.9 & movies$rating > 4.9
```

```
F <- movies$rating <= 4.9 & movies$rating >= 1

movies$Grade[A] <- c("A")
movies$Grade[B] <- c("B")
movies$Grade[C] <- c("C")
movies$Grade[D] <- c("D")
movies$Grade[F] <- c("F")



# Plots I create with the above conditions:

# It is important to examine in a general sense the
# production of films over the years

qplot(year, data = movies[interest, ], geom = "freqpoly",
  binwidth = 2, color = genre, xlab = "Year",
  ylab = "Number of Movies", main = "Production of
  Films Per Year per Genre")


# To better understand how ratings work, I use the
# created grade scale and modify the previous time
# series to show how the relative grades evolve
# over time

qplot(year, ..density.., data = movies[interest, ],
  geom = "freqpoly", binwidth = 2, color = Grade,
  xlab = "Year", main = "Time Series of Ratings")


# A further examination of the grade evolutions requires a
# facet by genre to show each genre's effect

qplot(year, ..density.., data = movies[interest, ],
  geom = "freqpoly", binwidth = 2, color = Grade,
  xlab = "Year", main = "Time Series of Ratings Faceted
  by genre") + facet_grid(. ~ genre)

# Using ddply, we can create a data frame that will permit
# us to examine the total votes per year

votes_ply <- ddply(movies, "year", summarise,
  total = sum(votes), mean = mean(votes))

qplot(year, total, data = votes_ply, xlab = "Year",
  ylab = "Total Votes", main = "Number of Votes per Year")
  + geom_smooth() + geom_vline(xintercept = 1990)
```